图书情报工作
第66卷 第11期 2022年6月

虚拟学术社区中融合用户动态兴趣与社交关系的学者推荐研究*

■ 顾佳云 熊回香 肖兵
华中师范大学信息管理学院 武汉 430079

摘 要：[目的/意义]考虑用户兴趣和社交关系两方面的动态变化，提出融合用户动态兴趣与社交关系的学者推荐模型。[方法/过程]首先，利用不同学科的期刊文献作为分类语料，基于 Labeled-LDA 模型对学者所发博文进行学科领域判别。然后，依据 KNN 算法对博文进行学科分类，接着利用学科兴趣变化速率改进时间因子，计算得到学者动态兴趣相似度；根据学者间链接的数量关系计算学者的 PageRank 值，结合学者所发博文的时间价值计算得到全局信任度。在学者评论、推荐交互行为中引入时间权重计算学者交互信任度，综合全局信任度和交互信任度得到学者的动态社交信任度。最后，融合兴趣相似度与信任度进行学者推荐。[结果/结论]虚拟学术社区中融合用户动态兴趣与社交关系的学者推荐模型从动态兴趣和动态社交关系两个视角出发，能够有效提高学者推荐的质量。

关键词：虚拟学术社区 动态兴趣 社交关系 学者推荐 Labeled-LDA 主题模型
分类号：G250
DOI：10.13266/j.issn.0252-3116.2022.11.012

1 引言

随着 Web2.0 时代的到来，社交网络飞速发展，根据第49次《中国互联网络发展状况统计报告》显示，截至2021年12月，我国网民规模达10.32亿，即时通信用户规模达10.07亿，社交网络已经成为人们通信交流的重要渠道[1]。随着社交网站的广泛应用，各种各样的在线虚拟社区为具有共同兴趣爱好和需求的用户提供了在线交流与互动的平台[2-3]。在虚拟社区中，用户可以通过频繁、双向的交流和合作，交换思想，实现用户之间的头脑风暴，融合个体用户的知识与智慧[4]。由于一般的综合性或大众化虚拟社区存在用户纷杂，成员流动性大，传播的信息内容、形式、质量不一等问题，一些专业虚拟社区如知识问答社区[5]、在线健康社区[6]、虚拟学术社区[7]等开始出现，它们将具有共同兴趣和特定领域知识的用户聚集在一起，交流专业知识和经验[8-9]。虚拟学术社区作为典型的专业虚拟社区，其以科研工作者为服务对象，支持科研人员知识交流、共享和维护社交关系[10]，研究显示虚拟学术社区能够有效提高学者的学术曝光率，促进新知识的产生和传播[11]。

然而，随着虚拟学术社区的快速发展，学术信息数量急速增长，为用户寻找与自己研究兴趣相投的学者带来了很多阻碍，很多研究者以学者兴趣特征为基础，融合多维度属性进行了学者推荐研究，其中部分研究者注意到了学者研究兴趣或社交关系的动态变化，但很少有研究同时考虑兴趣与社交关系两者的动态变化。因此，本文提出一种融合用户动态兴趣与社交关系的学者推荐模型，该模型利用学科兴趣变化速率改进时间因子，将时间因子引入学者的博文相似度和社交信任度计算中，同时考虑了学者在兴趣和社交行为两方面的动态变化，有效提高学者推荐的准确性。

2 相关研究

当前，基于用户兴趣挖掘推荐研究的视角已经从关注用户的静态兴趣向动态兴趣转变。国内围绕用户动态兴趣推荐的研究已在数字图书馆、微博社区、知识服务等领域有所探讨。潘家武[12]运用领域本体构建数字图书馆的动态用户兴趣模型，将实时获取的用户兴趣信息与领域本体库进行匹配修正，以匹配的方式满足用户个性化的需求；陶永才等[13]提出一种基于加权动态兴趣度（Weighted Dynamic Degree of Interest，WDDI）的微博个性化推荐模型，WDDI 模型在微博转发特征中引入时间因子，从而建立面向用户主题的个

* 本文系国家社会科学基金重大项目"新时代我国文献信息资源保障体系重构研究"（项目编号：19ZDA345）研究成果之一。
作者简介：顾佳云，硕士研究生；熊回香，教授，博士，博士生导师，通信作者，E-mail：hxxiong@mail.ccnu.edu.cn；肖兵，博士研究生。
收稿日期：2021-10-07 修回日期：2022-02-21 本文起止页码：110-120 本文责任编辑：徐健

体动态兴趣模型,另外通过用户与其关注用户的相似度和交互频率获取用户的群体动态兴趣,将用户个体兴趣与群体兴趣加权结合得到加权动态主题兴趣模型;应璇等[14]从用户知识范畴的获取、知识关联挖掘及动态知识演化角度测度用户兴趣随时间的推移而表现出的知识概念漂移及变化趋势。部分学者将动态兴趣与社交关系结合进行推荐,王占等[15]利用时间遗忘函数模拟用户的兴趣变化,然后将信任关系和兴趣变化融入到协同过滤推荐中;王维等[16]融入艾宾浩斯遗忘函数的 Pearson 相关系数以计算用户间的兴趣相似度,通过加权融合获取用户信任与用户兴趣间的关联关系,以获取更加准确的最近邻居。

在国外的动态兴趣研究中,H. Feng 等[17]基于时间加权关联规则的时间重叠群体算法来模拟用户的兴趣变化,克服了用户兴趣漂移导致推荐效果不好的问题。X. J. Liu 等[18]引入时间衰减函数来反映用户的兴趣变化,并使用改进的相似度模型进行推荐。S. Pariserum 等[19]在电子学习推荐系统中将数据流偏好划分到大小相等的窗口中,并随时间变化对用户兴趣评级,能够有效提高内容推荐的效率和准确性。也有少部分研究将兴趣的动态变化与社交关系相结合,如 C. Xu 等[20]将直接和间接信任关系、用户偏好、签到时间和地理位置融合到矩阵分解模型中进行兴趣推荐。

综上所述,目前学者们有关个性化推荐研究的视角已经从关注用户的静态属性转向动态属性,部分研究分别对用户兴趣和社交关系的动态性有所考虑,研究用户的兴趣迁移能够发现用户的近期兴趣,研究社交行为的动态性能够描述用户之间近期的社交关系,学者在研究兴趣和社交关系两个维度时考虑时间因素,兼顾了虚拟学术社区的学术性与社交性,但很少有研究同时考虑动态兴趣与动态社交关系。另外,目前关于动态性的研究,主要是利用遗忘函数计算时间权重,学科资源本身的老化速度带来的兴趣变化速率也较少考虑。因此,本文提出了一种融合用户动态兴趣与社交关系的学者推荐模型,该模型首先利用不同学科的期刊文献作为分类语料,基于 Labeled-LDA 模型对学者所发博文进行学科领域判别,然后利用 KNN 算法对博文进行学科分类;接着利用学科资源的老化速度来表示学者对某学科兴趣变化的速率,利用学科兴趣变化速率改进学者发表每篇博文的时间因子;将改进后的学科－时间因子引入学者博文主题矩阵得到学者动态兴趣矩阵,然后计算得到学者动态兴趣相似度。在学者信任度计算中,利用学者间链接的数量关系计算学者的 PageRank 值,结合学者所发博文的时间价值计算得到全局信任度;在学者评论、推荐交互行为中引入时间权重计算学者交互信任度,综合全局信任度和交互信任度得到学者的动态社交信任度。最后通过实验找到最合理的权重分配,融合兴趣相似度与社交信任度计算学者最终推荐评分,选取分数最高的 Top-K 学者进行推荐。

3 融合用户动态兴趣与社交关系的学者推荐模型

随着时间的迁移,虚拟学术社区中产生了大量的用户社交行为和学术信息,考虑学者在不同时间的兴趣变化和交互行为,可以更准确及时地反映学者当前的兴趣和社交关系,提高学者推荐效果。本文提出的融合用户动态兴趣与社交关系的虚拟学术社区中学者推荐模型包括 4 个模块,分别是数据采集与预处理模块、学者动态兴趣模块、学者动态社交模块和学者推荐模块。如图 1 所示:

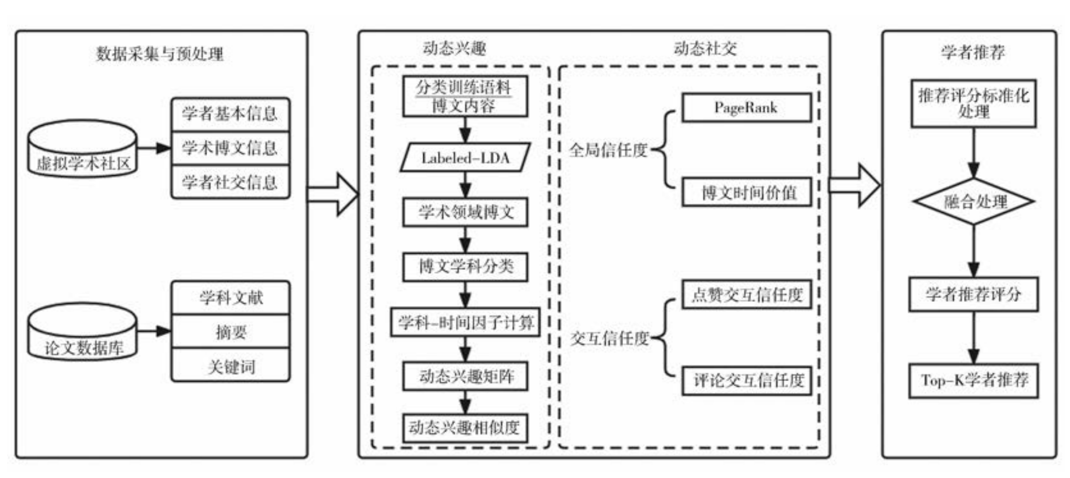


图 1 融合用户动态兴趣与社交关系的学者推荐模型

（1）数据采集与预处理模块。首先采集虚拟学术社区中相关学者数据，并采集相关学科的期刊文献数据作为博文分类语料，然后，对数据进行去重、缺失值处理得到学者基本信息、博文信息和社交信息，以及期刊文献的摘要、关键词等。

（2）学者动态兴趣模块。利用不同学科的期刊文献作为分类语料，基于 Labeled-LDA 模型对学者所发博文进行学科领域判别，然后利用 KNN 算法对博文进行学科分类；学者对某学科的研究兴趣可以表现为对该学科资源的利用情况，而学科资源的利用情况可以通过该学科的资源老化情况来衡量，因此，可以利用学科资源的老化速度来表示学者对该学科兴趣变化的速率，然后利用学科兴趣变化速率改进学者发表每篇博文的时间因子；将改进后的学科－时间因子引入学者博文主题矩阵，加权平均得到学者动态兴趣特征向量，最后利用余弦相似度计算用户动态兴趣相似度。

（3）学者动态社交模块。学者动态社交通过社交信任度来表示，在信任度计算中，利用学者间链接的数量关系计算学者的 PageRank 值，结合学者所发博文的时间价值计算得到全局信任度；在学者评论、推荐交互行为中引入时间权重计算学者交互信任度，综合全局信任度和交互信任度得到学者的动态社交信任度。

（4）学者推荐模块。通过实验找到最合理的权重分配，融合学者动态兴趣相似度和动态社交信任度进行推荐评分，选取分数最高的 Top-K 学者推荐给目标用户。

3.1 基于用户动态兴趣的相似度计算

3.1.1 学科分类语料构建

根据邱均平等[21]的研究，参照图书情报领域学者跨学科研究的前 20 个学科领域，本文选择了图书情报学、计算机科学、新闻学与传播学、高等教育学、生物信息学、管理科学与工程 6 个学科领域的学者作为研究对象。不同于文献数据库中学者的期刊论文，学者在虚拟学术社区中生成的内容可能存在生活分享、话题讨论等非学术性内容，同一个学者也可能发表涉及不同学科的内容，不能仅通过学者所属学科类别判定博文学科，所以需要对博文进行学术信息识别和学科分类，本文从 CNKI 中采集选定学科的核心期刊文献摘要和关键词等信息作为分类语料，将期刊文献的所属学科作为分类标签。

3.1.2 主题特征提取

本文利用 Labeled-LDA 模型同时对有标记的期刊文献和未标记的博文进行主题特征提取，得到博文文档向量和期刊文档向量，Labeled-LDA 模型是 D. Ramage 等[22]于 2009 年在 LDA 模型的基础上提出的一种有监督的主题模型，其主要用于对有标签的文档进行建模，与 LDA 模型相比，Labeled-LDA 模型多出一层为每篇文档附加的类别标签 Λ，运用某篇文章是否属于一个标签类别（Λ 值）来约束文档的主题概率分布（θ 值）。在选择一个词的时候，LDA 模型是在所有的主题上选择该词，而 Labeled-LDA 模型则是只从文档相关的标签所对应的主题中去选择，避免了词在所有主题上的分配，将词的主题范围限定在所属文档标记的主题之内，很好地利用了人工标记的主题信息[23]，可以避免文档在不对应的分类上进行强制分配的缺陷，如图 2 所示：

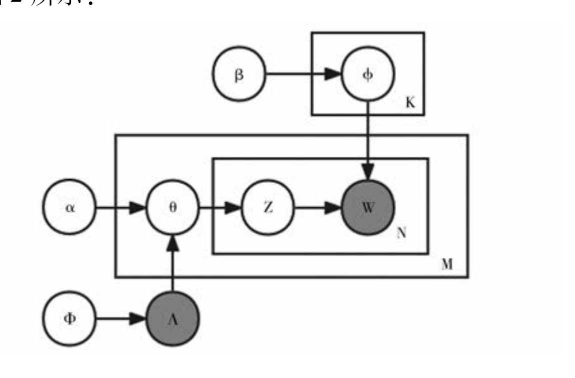


图 2 Labeled-LDA 模型

其中，M 代表文档集合，K 代表主题总数，N 是每篇文章中含有的总词数，隐变量 Z 表示某一个主题，W 是文本的单词，θ 为文本－主题概率分布，φ 为主题－语词概率分布，参数 α 和 β 分别是 θ 和 φ 的超参数，Λ 是某篇文章的学科标签类别。

3.1.3 博文学科分类

本文利用平均主题相关度阈值规则判别博文是否属于选定学科领域，将 Labeled-LDA 主题模型训练得到的博文文档向量和期刊文档向量分别记为 θ_i（$i=1,2,\cdots,N$）（N 为博文数）和 θ_j^t（$j=1,2,\cdots,M$），（$t=1,2,\cdots,T$）（M 为某学科的期刊文献数，T 为学科数），计算博文 i 与学科领域 t 的平均主题相关度[24]，如公式（1）所示：

$$S_{ij} = \sum_{j=1}^{M} \frac{1-\sqrt{2\times D_{JS}(i,j)}}{M} \qquad \text{公式（1）}$$

其中，$D_{JS}(i,j)$ 是 θ_i 和 θ_j^t 的 Jensen-Shannon 散度，M 为学科 t 的期刊文献数，$0<S_{ij}<1$，S_{ij} 的值越高说明博文 i 与这个学科领域平均主题越相似，由此可以对学者所发博文进行学科领域判别。

在完成学科领域博文选取后，为了提升分类准确性，本文利用 KNN 算法对博文进行学科分类。

3.1.4 学者的动态兴趣相似度计算

学者研究兴趣的动态变化与人类大脑的遗忘过程较为相似，因此根据学者不同阶段发表的博文，采用模拟遗忘函数来模拟学者的兴趣变化。

引入时间因子 $W(u,i)$，作为学者不同时期研究兴趣的权重，如公式(2)所示：

$$W(u,i)=1/e^{a\sqrt{\frac{t_{last}-t_i}{t_{last}-t_{first}}}} \quad \text{公式(2)}$$

其中 t_i 为与当前待加权博文 i 对应的时间($t_{first}<t_i<t_{last}$)，t_{last} 表示学者最新一篇博文对应的时间，t_{first} 表示学者最早一篇博文对应的时间，t_i 的值越大，表示 $W(u,i)$ 的值越大($e^{-1}\leqslant W(u,i)\leqslant 1$)，说明该博文的发表时间越近，分配的权重越大，越能代表该学者当前的兴趣[25]。a 为学科兴趣变化速率，对于不同学科领域而言，博文老化的速度是不相同的[26]，a 的值越小，表示学者对该学科资源兴趣变化的速度越慢，相较于其他学科同一时间发布的博文，该学科博文分配的权重越大，越能代表该学者当前的兴趣。

网络信息的老化符合负指数函数如公式(3)所示：

$$C(t_t,t_i)=e^{-a(t_t-t_i)} \quad \text{公式(3)}$$

其中 t_i 表示信息发布的时间，t_t 表示当前的时间，$C(t_t,t_i)$ 表示信息在 t_t 时刻的影响力大小，a 代表的是信息的老化率系数，化简公式(3)计算老化系数 a[27]，如公式(4)所示：

$$a=-\frac{ln(C(t_t,t_i))}{t_t-t_i} \quad \text{公式(4)}$$

当 $C(t_t,t_i)=1/2$ 时，t_t-t_i 实际是信息的半衰期，记为T。根据半衰期的定义，信息的半衰期是指信息自被发布的时刻开始到信息的利用率下降到一半的时刻之间的时间段，即信息的影响力减半的时间，这里用某学科所有博文 T 的平均值表示该学科的博文半衰期 $\bar{T}$，如公式(5)所示：

$$\bar{T}=\frac{\sum_o^n T(i)}{n} \quad \text{公式(5)}$$

其中 n 为该学科博文数，$T(i)$ 为博文 i 的半衰期，即博文从发布到评论数达到一半的时间，最后根据公式(6)计算出某学科博文的老化系数 a。

$$a=-\frac{ln(0.5)}{\bar{T}} \quad \text{公式(6)}$$

学者动态兴趣向量可以表示为 User $=(F_1,F_2,\cdots F_n)$，其中 $F_j(j=1,2,\cdots n)$ 对应为学者在第 j 个主题上的动态向量值，如公式(7)所示：

$$F_j=\sum_0^q x_{ij}W(u,i)/q \quad \text{公式(7)}$$

其中 x_{ij} 表示学者第 $i(i=1,2,\cdots,q)$ 篇博文在主题 j 下的概率，$W(u,i)$ 表示学者第 $i(i=1,2,\cdots,q)$ 篇博文的时间权重。

最后利用余弦相似度计算学者动态兴趣相似度，如公式(8)所示：

$$sim(u,v)=\cos(u,v)=\frac{\sum_1^n u_j\times v_j}{\sqrt{\sum_1^n u_j^2}\times\sqrt{\sum_1^n v_j^2}} \quad \text{公式(8)}$$

式中，学者 u 和学者 v 分别用($u_1,u_2,\cdots,u_n$)和($v_1,v_2,\cdots,v_n$)表示，n 表示所有主题数；u_j 和 v_j 分别表示学者 u 与学者 v 在第 j 个主题下的动态向量值。

3.2 基于用户动态社交关系的信任度计算

社交网络中用户间的关系可以划分为全局关系和交互关系，全局关系表示用户在全局网络中的信誉和影响力，用全局信任度表示，交互关系表示两两学者之间的互动行为，用交互信任度表示[28]。

3.2.1 学者全局信任度计算

学者的全局信任度可以通过学者在全局信任网络中的声誉和影响力来衡量，比较具有代表性的算法是 PageRank 算法，PageRank 算法是谷歌在搜索引擎结果中用于对网站进行排名的算法，其核心思想是网页的重要性通过其他网页对其链接的数量衡量[29]，然而 PageRank 仅考虑了网站网页间链接的数量关系，并不对其本身价值进行分析[30]。所以本文结合学者好友关系和所发博文的价值进行分析，一方面发掘学者在整个虚拟学术社区中的影响力，另一方面衡量学者所发布博文的价值。

本文首先借鉴 PageRank 的算法思想，将该算法中的网页链接关系换成学者链接关系计算得到学者的 PageRank 值，然后计算学者所发博文的价值。学术文献的价值可以用期刊级别、发表时间、被引用情况等来衡量[31]，而网络信息更注重时效性。本文通过计算博文的时间价值来体现博文的使用价值[32]。将学者的 PageRank 值和学者所发博文的价值相结合，计算得到学者全局信任度，如公式(9)所示：

$$\mathrm{D(u)}=\mathrm{PR(u)}+\sum_1^q 2^{\bar{T}-t_i} \quad \text{公式(9)}$$

其中 PR(u) 为学者 u 的 PageRank 值，$\bar{T}$ 为博文学科半衰期，由公式(5)计算得到，t_i 表示博文 i 的产出年龄，q 表示学者 u 产出的博文数，$2^{\bar{T}-t_i}$ 表示学者 u 第 i 篇博文的价值[33]。

3.2.2 学者交互信任度计算

不同于全局信任关系，学者交互信任度是要计算两两学者之间的信任度，可以通过学者之间的历史交

互行为计算学者交互信任度，同时学者之间的交互行为是动态变化的，越近的交互行为信任权重越高[34]。学者在虚拟学术社区中的交互行为主要表现为对学者所发博文的点赞和评论。

由于无法准确获取用户的点赞时间，所以点赞交互行为 F(u,v) 没有考虑时间权重，如公式(10)所示：

$$F(u,v) = \frac{B(u,v)}{q_v} \qquad \text{公式(10)}$$

其中，$B(u,v)$ 表示学者 v 对学者 u 的点赞数，q_v 表示学者 v 的点赞行为数。

学者 v 对学者 u 的评论行为产生的交互信任度 G(u,v) 如公式(11)所示：

$$G(u,v) = \sum_1^n L_j \times w_j \qquad \text{公式(11)}$$

其中 L_j 表示第 j 年度的评论交互信任度，如公式(12)所示，$times_{vj}$ 表示第 j 年度学者 v 的评论行为数，$times_{uvj}$ 表示第 j 年度用户 v 对用户 u 的评论次数，n 表示评论交互持续时间，w_j 表示第 j 年度的评论交互时间权重，如公式(13)所示，t_t 表示当前年份，t_f 表示评论交互的年份，评论交互的时间越近，则权重越大。

$$L_j = \frac{times_{uvj}}{times_{vj}} \qquad \text{公式(12)}$$

$$w_j = e^{\frac{1}{t_t - t_f}} \qquad \text{公式(13)}$$

综合学者 v 对学者 u 的点赞交互数和评论交互数计算得到学者 v 对学者 u 的交互信任度 T(u,v)，如公式(14)所示：

$$T(u,v) = \omega \times F(u,v) + (1-\omega) \times \sum_1^q G(u,v) \qquad \text{公式(14)}$$

其中，$0 \leqslant \omega \leqslant 1$，F(u,v) 表示学者 v 对学者 u 的点赞交互信任度，G(u,v) 表示学者 v 对学者 u 的评论交互信任度。

3.2.3 学者动态社交信任度融合计算

学者的全局信任度反映了学者在整个虚拟学术社区中的影响力和地位，学者的交互信任度反映了学者网络节点间的交互信任度。对全局信任度 D(u) 和交互信任度 T(u) 计算之后，将两个数值进行线性加权得到的学者动态社交信任度 Q(u)，如公式(15)所示：

$$Q(u) = \beta \times D(u) + (1-\beta) \times T(u,v) \qquad \text{公式(15)}$$

在公式(15)中，β 是全局信任度 D(u) 和交互信任度 T(u) 的融合参数。如果 $\beta > 0.5$，表明全局信任度较交互信任度更为重要；如果 $\beta < 0.5$，表明交互信任度更为重要，本文设定参数 $\beta = 0.5$，即认为全局信任度和交互信任度同样重要。

3.3 融合用户动态兴趣相似度与社交信任度的学者推荐

学者动态兴趣相似度反映了学者动态兴趣倾向，学者的动态信任度反映了学者在整个虚拟学术社区中的影响力和地位以及学者之间点对点的交互关系。对学者动态兴趣相似度 sim(u,v) 和动态信任度 $Q(u)$ 计算之后，将两个数值进行线性加权得到的综合数值，即待推荐学者的推荐评分 S_{user}，然后根据推荐评分向目标用户进行 Top-K 学者推荐。最终推荐评分可表示如下：

$$S_{user} = \gamma \times sim(u,v) + (1-\gamma) \times Q(u) \qquad \text{公式(16)}$$

公式(16)中，γ 是学者动态兴趣相似度和动态信任度的融合参数，如果 $\gamma > 0.5$，表明学者动态兴趣相似度较动态信任度更为重要；如果 $\gamma < 0.5$，表明动态信任度更为重要，这里 γ 根据实验情况设定。

4 实证分析

4.1 数据收集与预处理

为了对该推荐模型进行验证，本文首先确定了 6 个选定学科领域的核心期刊目录，然后从 CNKI 中采集了这些期刊 2015 年至今的所有文献信息，最终采集到 117 029 篇文献信息，6 个学科领域的核心期刊目录及文献数见表 1。

另外，利用八爪鱼爬虫工具获取科学网博客[35]中 6 个学科领域学者的博文及其社交信息，去除博文数量或好友数量为 0 以及隐私设置不可见的学者，采集了 217 名学者 2015－2021 年间所有的博文信息数据和社交数据，包括博文标题、时间、正文、评论、推荐、推荐时间以及好友列表。在剔除掉不完整、无效的博文之后，最终得到了 217 名学者的 24 081 条博文以及博文推荐、评论和所有好友。其中，学者博文数据、学者评论数据、学者好友数据分别见表 2、表 3、表 4。

4.2 学者动态兴趣相似度计算

4.2.1 文本主题特征提取

研究首先从 117 029 篇期刊文献中随机抽取每个学科领域 5 000 篇文献作为训练语料，然后剔除博文数据集中的一些特殊字符后，利用 Python 中 NLPIR 包，并结合停用词表与用户自定义词典，对学者博文和期刊文献摘要进行分词。将 30 000 篇带有 6 个不同学科标签的期刊文献和 24 081 篇没有标签的博文加入 Labeled-LDA 模型训练，主题标识即学科标签，得到主题－词项概率分布以及文本－主题概率分布，各主题下概率最大的 10 个词汇见表 5。然后通过公式(1)计算博文与各个学科领域的平均主题相似度，阈值设定

顾佳云，熊回香，肖兵.虚拟学术社区中融合用户动态兴趣与社交关系的学者推荐研究[J].图书情报工作,2022,66(11):110-120.

为0.5,最终得到22 938篇属于6个学科领域的博文,为了提升分类准确性,本文利用KNN算法进一步确定博文所属的学科类别。

表1　核心期刊目录及文献数

单位/篇

图书情报学	计算机科学	新闻学与传播学	高等教育学	生物学	管理科学与工程
大学图书馆学报,792	计算机辅助设计与图形学学报,1 702	编辑学报,1 867	大学教育科学,1 001	生物工程学报,1 422	管理工程学报,804
情报科学,2 356	计算机工程,4 136	编辑之友,1 641	高等工程教育研究,1 495	生物化学与生物物理进展,811	管理科学,482
情报理论与实践,2 336	计算机工程与应用,5 976	出版发行研究,2 129	高等教育研究,1 193	生物技术,1 393	管理科学学报,667
情报学报,872	计算机集成制造系统,2 050	当代传播,1 174	高教探索,2 435	生物技术进展,606	管理评论,1 977
情报杂志,2 607	计算机科学,6 249	国际新闻界,843	江苏高教,1 822	生物技术通报,2 447	管理世界,1 162
情报资料工作,672	计算机学报,1 024	科技与出版,2 487	现代大学教育,598	生物技术通讯,996	管理学报,1 466
数据分析与知识发现,985	计算机研究与发展,1 613	现代出版,875	中国高等教育,3 164	生物信息学,252	南开管理评论,738
图书情报工作,3 171	计算机应用,5 155	现代传播,2 729	中国高教研究,1 599	生物学杂志,1 181	系统工程,1 260
图书情报知识,585	计算机应用研究,5 582	新闻大学,781		生物医学工程学杂志,1 139	系统工程理论与实践,2 004
图书与情报,815	软件学报,1 541	新闻记者,1 029		中国生物工程杂志,1 238	系统工程学报,492
现代情报,2 062	中国图象图形学报,1 331	新闻与传播研究,705			系统管理学报,864
中国图书馆学报,346		中国编辑,1 402			运筹与管理,2 017
		中国出版,3 244			中国管理科学,1 960
		中国科技期刊研究,1 480			

表2　学者博文数据

学者ID	标题	时间	博文内容	点赞学者ID
3410526	入选2022年度浙江省"尖兵""领雁"研发攻关计划项目会评专家名单	2021/9/29	人工智能组,项目经费都是500W/项,浙江省2022年度"尖兵""领雁"研发攻关计划项目……	750818、107667
	……	……	……	……
	【围城围谁的城】——兼答青年博士求职FAQ(续)	2019/12/14	【围城围谁的城】——兼答青年博士求职FAQ(续)自老刘上次发文【围城围谁的城】——兼答青年博士求职……	325385、425437、1213429……
……	……	……	……	……
542	关于科(医)学,来点思考	2020/3/10	讨论一些热门话题如中医西医、转基因非转基因,会让和谐相处的人们站到不同队列中,科学知识具有社会和文化属性……	5889、1536597、107667……
	……	……	……	……
	五部门关于科技期刊的"意见",从南京大学开始试点如何?	2015/11/26	五部门关于科技期刊的"意见",从南京大学开始试点如何?2015年11月4日,中国科学技术协会、教育部……	769161、41701、1458267……

表3　学者评论数据

学者ID	评论学者ID	评论时间
3410526	1200905	2021/11/23
	……	……
	1213429	2019/12/15
1835014	3436271	2020/12/15
	……	……
	561693	2015/5/22
……	……	……
542	107667	2021/8/16
	……	……
	561693	2015/1/19

表4　学者好友数据

学者ID	好友ID
3410526	3497110、3493616、3492623、3469996、3466976……
1835014	3451787、1898783、2381229、475、1943390……
1037866	425437、38450、1350441、1834487、2460165……
1213429	60980、2649160、669170、1200905、729911……
……	……
542	3408518、3360562、3354122、3316859、3260634……

表 5 主题 - 词项概率分布(与主题相关的 10 个高概率词)

Topic	主题标识	词项(与主题相关的 10 个高概率词)
0	新闻学与传播学	出版、期刊、中国、发展、传播、研究、我国、新闻、内容、科技期刊
1	计算机科学	算法、方法、图像、模型、提高、数据、利用、特征、性能、实验结果
2	图书情报学	研究、分析、图书馆、方法、我国、领域、服务、发展、数据、构建
3	高等教育学	发展、高校、学生、大学、建设、教育、我国、研究、高等教育、教师
4	管理科学与工程	企业、研究、影响、本文、市场、关系、模型、中国、产品、发现
5	生物学	基因、细胞、表达、研究、蛋白、检测、分析、方法、利用、构建
6	common_topic	研究、方法、模型、用户、分析、影响、信息、数据、本文、构建

4.2.2 学者的动态兴趣特征提取

通过公式(2)计算学者动态兴趣时间因子,计算各个学科领域所有博文从发布到评论数达到一半的时间,然后利用公式(6)计算出某学科博文的老化系数 a。通过公式(7)计算学者动态兴趣向量,部分结果如表 6 所示:

表 6 学者动态兴趣向量

学者 ID	主题 0	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6
3410526	0.098 54	0.193 45	0.044 72	0.022 47	0.351 44	0.021 28	0.005 94
3389532	0.187 37	0.254 77	0.155 62	0.013 58	0.100 49	0.011 73	0.021 70
3388899	0.166 85	0.266 49	0.214 25	0.013 09	0.083 73	0.013 49	0.017 55
3334560	0.066 80	0.255 16	0.156 41	0.039 88	0.019 52	0.004 70	0.317 12
3316383	0.114 68	0.240 16	0.059 93	0.109 36	0.171 58	0.057 63	0.030 92
……	……	……	……	……	……	……	……
542	0.262 74	0.185 17	0.022 72	0.042 25	0.100 44	0.045 15	0.035 48

4.2.3 学者动态兴趣相似度计算

基于学者动态兴趣向量,利用公式(8)计算学者博文相似度,部分结果如表 7 所示:

表 7 动态兴趣相似度矩阵

学者 ID	3410526	3389532	3388899	3334560	3316383	……	542
3410526	1	0.722 76	0.661 92	0.393 37	0.862 89	……	0.696 68
3389532	0.722 76	1	0.986 84	0.682 90	0.879 78	……	0.882 49
3388899	0.661 92	0.986 84	1	0.695 75	0.835 86	……	0.803 68
3334560	0.393 37	0.682 90	0.695 75	1	0.625 43	……	0.545 65
3316383	0.862 89	0.879 78	0.835 86	0.625 43	1	……	0.849 15
……	……	……	……	……	……	……	……
542	0.696 68	0.882 49	0.803 68	0.545 65	0.849 15	……	1

4.3 学者动态社交信任度计算

4.3.1 全局信任度计算

利用公式(9)计算学者全局信任度,其中 PageRank 值通过 Python 计算得到,同时考虑学者所发博文本身的时间价值,部分结果如表 8 所示:

表 8 学者全局信任度

学者 ID	PageRank	时间价值	全局信任度
3410526	0.405 56	0.006 84	0.412 40
3389532	0.061 94	0.031 24	0.093 18
3388899	0.011 30	0.060 12	0.071 42
3334560	0.101 39	0.438 17	0.539 57
3316383	0.111 03	0.174 59	0.285 62
……	……	……	……
542	0.039 42	0.007 83	0.047 25

4.3.2 交互信任度计算

学者动态交互信任度由学者的点赞交互信任度和评论交互信任度构成。点赞交互信任度利用公式(10)计算得到,评论交互信任度利用公式(11)计算得到,两者结合得到学者动态交互信任度,部分结果见表 9。

4.3.3 动态社交信任度融合计算

利用公式(15)融合全局信任度和交互信任度计算学者动态社交信任度,结果见表 10。

4.4 学者推荐评分融合计算

利用公式(16)融合动态兴趣相似度和动态社交信任度,得到学者最终推荐评分,融合参数 γ 根据实验

表 9　学者交互信任度

学者 ID	3410526	3389532	3388899	3334560	3316383	……	542
3410526	1	0	0	0	0	……	0
3389532	0	1	0.037 47	0	0	……	0
3388899	0	0	1	0	0	……	0
3334560	0	0	0	1	0	……	0
3316383	0.291 06	0	0.027 78	0	1	……	0
……	……	……	……	……	……	……	……
542	0	0	0	0	0	……	1

表 10　学者动态社交信任度

学者 ID	3410526	3389532	3388899	3334560	3316383	……	542
3410526	0.251 51	0.206 20	0.206 20	0.206 20	0.206 20	……	0.206 20
3389532	0.046 59	0.046 59	0.065 32	0.046 59	0.046 59	……	0.046 59
3388899	0.035 71	0.035 71	0.802 92	0.035 71	0.035 71	……	0.035 71
3334560	0.269 78	0.269 78	0.269 78	0.269 78	0.269 78	……	0.269 78
3316383	0.288 34	0.14 281	0.156 70	0.142 81	1.661 19	……	0.142 81
……	……	……	……	……	……	……	……
542	0.023 63	0.023 63	0.023 63	0.023 63	0.023 63	……	0.391 06

情况设定，在好友推荐结果评价中，当前较为常用的评价指标为准确率(Precision)、召回率(Recall)以及综合准确率和召回率的 F1-measure[36-37]。其公式如下：

$$准确率 = \frac{推荐出的已经成为好友的学者数}{推荐正确的学者数 + 推荐错误的学者数} \quad 公式(17)$$

$$召回率 = \frac{推荐出的已经成为好友的学者数}{推荐正确的学者数 + 应该被推荐但没有被推荐的学者数} \quad 公式(18)$$

$$F1-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad 公式(19)$$

实验比较了 γ 为 0.5、0.6、0.7 的三种情况，当推荐学者数量少于 15 时，γ=0.6 效果最好，推荐学者数量超过 15 时，γ=0.5 效果最好，结果见图 3-图 5，本文最终选择 γ=0.6 进行评分融合，为便于对后续结果分析，对角线上的结果一律设为 -1，结果见表 11。

5　推荐结果分析

融合用户动态兴趣相似度和社交关系信任度，最终得到推荐结果，部分数据见表 12。

随机选取学者 3410526 推荐的前 5 名学者为例进行验证，可以看出本文提出的推荐模型可以得到较好的推荐结果，学者 3410526 为教授、研究生导师，其所属的研究领域为信息科学大类。

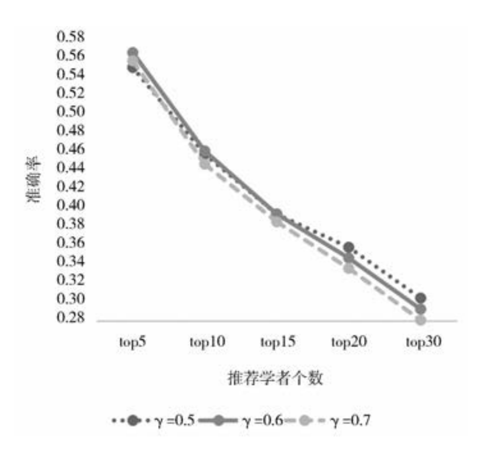


图 3　融合参数 γ 对准确率的影响

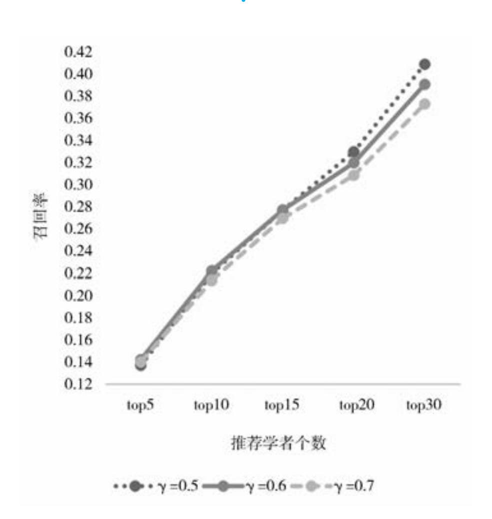


图 4　融合参数 γ 对召回率的影响

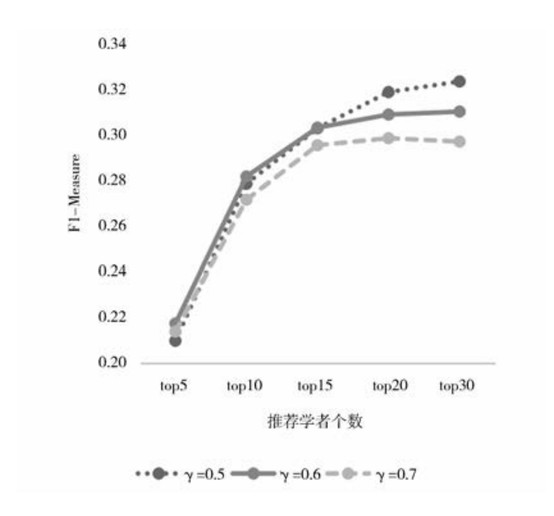


图 5 融合参数 γ 对 F1-Measure 的影响

从学者研究兴趣分析，学者 3410526 发表的博文较多关注科学研究、研究生教育、博士就业相关内容。学者 1557 发表过很多科学研究方面的博文，例如“像打理花园一样培育科研”“科技期刊在数字化时代的两难处境”“文献计量学与科技期刊研究”等博文；学者 425437 发表过很多博士就业相关的博文，例如“入职高校的博士生，要不要立马换方向”“一位博士生对就业选择的困惑”等博文；学者 2999994 发表过很多研究生教育有关的博文，例如标题为“导师和‘只想拿个学位’的研究生如何成为命运共同体?”“如何让导师和研究生成为命运共同体?”等博文；学者 522469 发表了很多研究生教育和科学研究相关的博文，例如“突然想把课程内容弄的好一点”“科研选大问题还是选小问题(科研生态中的岛屿效应)”“一个科研领域的兴衰——要不要坚守?”等博文；学者 3316383 发表了很多科学研究、论文撰写方面的博文，例如“如何在写文章时使用学术资源?”“使用文本挖掘工具对专有数据库的数据进行热门话题研究”等博文，这些都与目标学者 3410526 的研究兴趣有很高的相关性。

表 11 学者最终推荐评分

学者 ID	3410526	3389532	3388899	3334560	3316383	……	542
3410526	-1	0.516 14	0.479 63	0.318 50	0.600 22	……	0.500 49
3389532	0.452 29	-1	0.618 23	0.428 38	0.546 51	……	0.548 13
3388899	0.411 44	0.606 39	-1	0.431 73	0.515 80	……	0.496 49
3334560	0.343 93	0.517 66	0.525 36	-1	0.483 17	……	0.435 30
3316383	0.633 07	0.584 99	0.564 19	0.432 38	-1	……	0.566 62
……	……	……	……	……	……	……	……
542	0.427 46	0.538 94	0.491 66	0.336 84	0.518 94	……	-1

表 12 学者推荐结果

学者 ID	推荐结果
3410526	1557、425437、2999994、522469、3316383、215715、1213429、2903646、401512、64000、57940、3075、1256692、359436、652078……
3389532	1557、57940、3075、2374、287179、1968、826653、1750、94143、290937、951291、554179、496649、404304、215715……
3388899	1557、3075、2374、826653、290937、554179、1968、951291、496649、1750、287179、2322490、340399、94143、3389532……
3334560	1557、3075、290052、118204、404304、571917、508476、1835014、2577109、554179、787764、656335、1125809、707141、2636671……
3316383	1557、213646、3075、554179、94143、287179、215715、1792012、45134、1750、1213429、2374、57940、472757、425437……
……	……
542	1557、3075、213646、69474、61772、287179、290052、1968、215715、404304、45134、71721、576665、3503、94143……

从学者地位和活跃度分析，5 位推荐学者均为高校教授，在所属专业领域具备一定的声望，产出博文具有较高的学术价值，在虚拟学术社区中比较活跃。

为了进一步验证本文所提出的推荐模型在学者推荐方面的有效性，参照好友推荐的经典方法，本文比较了基于兴趣推荐、基于兴趣与社交关系推荐和融合动态兴趣与社交关系推荐的三种学者推荐算法的性能，其中基于兴趣的推荐和基于兴趣与社交关系的推荐没有考虑用户兴趣或社交行为的动态变化。由于部分学者的好友数过少，难以对推荐结果进行评价，所以在结果评价时设定好友数阈值为 10，对高于阈值的学者进行评价，结果如图 6 - 图 8。可以看出，本文方法推荐效果要明显优于基于兴趣和基于兴趣与社交关系的推荐方法。

6 结语

虚拟学术社区以科研工作者为服务对象，支持科研人员知识交流、共享和维护社交关系，极大地拓宽了传统意义上学术交流的途径，促进了非正式学术交流的发展，对新知识的创造、传播和不同学科知识的交叉渗透产生了重要影响。然而，虚拟学术社区中的注册用户和学术信息的海量增长使得如何帮助用户快速寻找学术水平较高、兴趣相投的学者成为目前迫切需要

顾佳云，熊回香，肖兵. 虚拟学术社区中融合用户动态兴趣与社交关系的学者推荐研究[J]. 图书情报工作,2022,66(11):110－120.

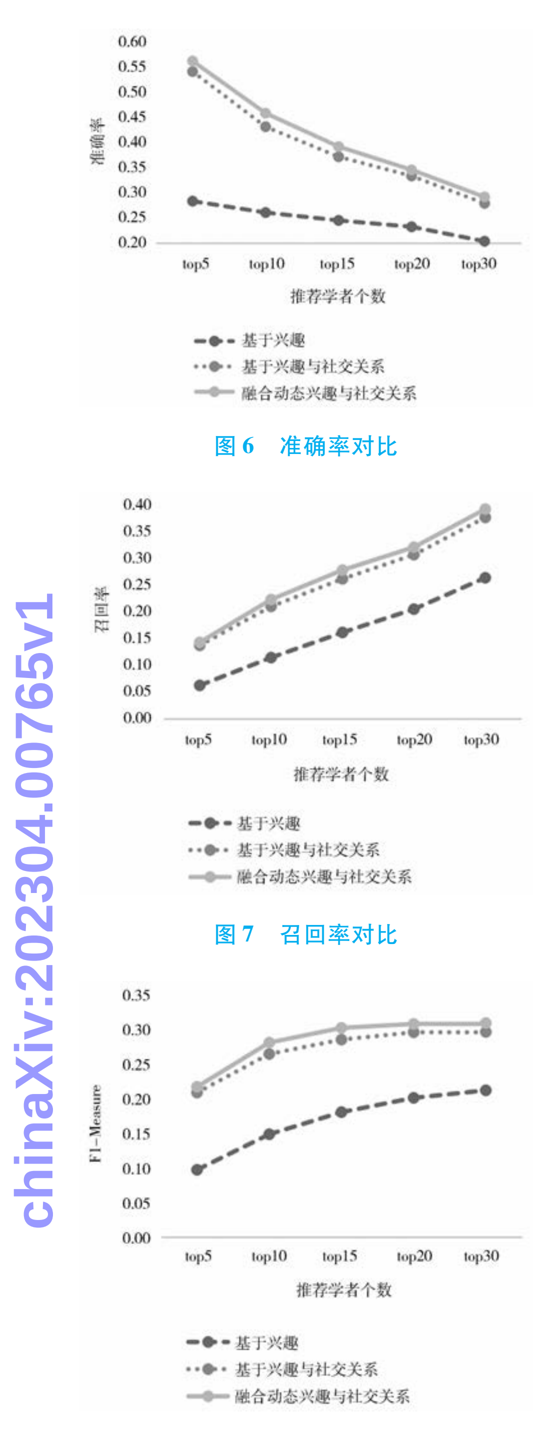


图 6　准确率对比

图 7　召回率对比

图 8　F1-measure 对比

处理的关键问题。学者推荐是处理虚拟学术社区中信息过载问题的首要选择，传统的学者推荐方法已经开始考虑学者研究兴趣的动态变化，但仅从学者的科研产出考虑兴趣的迁移，忽视了学者动态社交行为产生的影响。另外，目前关于动态性的研究，主要是利用模拟遗忘函数对学者不同时期的兴趣加权，很少从学科角度考虑学者群体兴趣的偏移。本文提出了一种融合动态兴趣与社交关系的学者推荐模型，在学者动态兴趣和动态社交信任度特征提取过程中，利用学科兴趣变化速率改进时间因子，兼顾了虚拟学术社区的学术性、社交性和不同学科特性。结果显示，虚拟学术社区中融合用户动态兴趣与动态社交关系的学者推荐模型在研究兴趣和社交关系两方面考虑时间因素，有效提高了学者推荐的质量。本文研究也存在一些不足之处，在学者评论交互信任度计算时未考虑学者回复行为产生的交互信任度，另外也未考虑学者之间的间接交互行为，这些问题有待在后续的研究中进一步完善。

参考文献：

[1] 中国互联网信息中心. 第 49 次中国互联网络发展状况统计报告[EB/OL]. [2022－02－20]. http://cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202202/P020220318335949959545.pdf.

[2] 王刚，郭雪梅. 社交网络环境下基于用户行为分析的个性化推荐服务研究[J]. 情报理论与实践，2018，41(8):102－107.

[3] 刘征驰，田小芳，石庆书. 网络虚拟社区知识分享治理机制[J]. 管理学报，2015，12(9):1394－1401.

[4] 邓卫华，易明，王伟军. 虚拟社区中基于 Tag 的知识协同机制——基于豆瓣网社区的案例研究[J]. 管理学报，2012，9(8):1203－1210.

[5] 齐云飞，赵宇翔，朱庆华. 在线问答社区中参与者知识行为研究综述[J]. 图书情报知识，2018(3):103－112.

[6] 席海涛，聂文博，李闺臣，等. 在线健康社区用户交互的研究现状与进展[J]. 情报科学，2021，39(4):186－193.

[7] 孙思阳，张海涛，任亮，等. 虚拟学术社区用户知识交流行为研究综述[J]. 情报科学，2019，37(1):171－176.

[8] HUNG S，LAI H，CHOU Y. Knowledge-sharing intention in professional virtual communities: a comparison between posters and lurkers[J]. Journal of the Association for Information Science & Technology，2015，66(12):2494－2510.

[9] KANG H，HAN J，KWON G. An integrated success factor model of professional virtual communities: incorporation of the operators，members，and life cycle perspectives[J]. International journal of human-computer interaction，2019，35(14):1312－1330.

[10] 苏晓兰. 学术虚拟社区用户交流模式研究[D]. 上海：华东师范大学，2015.

[11] NANDEZ G，BORREGO A. Use of social networks for academic purposes: a case study[J]. The electronic library，2013，31(6):781－791.

[12] 潘家武. 基于领域本体的数字图书馆动态用户兴趣模型的构建[J]. 图书情报工作，2010，54(8):64－67.

[13] 陶永才，何宗真，石磊，等. 基于加权动态兴趣度的微博个性化推荐[J]. 计算机应用，2014，34(12):3491－3496.

[14] 应璇，孙济庆. 面向知识服务的用户兴趣动态知识空间模型研究[J]. 情报学报，2017，36(2):206－212.

[15] 王占，林岩. 基于信任与用户兴趣变化的协同过滤方法研究[J]. 情报学报，2017，36(2): 197－205.

[16] 王维，高岭，高全力. 融合用户信任和用户兴趣漂移的协同过滤算法[J]. 微电子学与计算机，2019，36(7):103－108.

[17] FENG H，TIAN J，WANG H J，et al. Personalized recommendations based on time-weighted overlapping community detection[J]. Information & management，2015，52(7):789－800.

[18] LIU X J. An improved clustering-based collaborative filtering recommendation algorithm[J]. Cluster computing, 2017, 20(2): 1281－1288.
[19] PARISERUM S, SANNASI G, ARPUTHARAJ K. An intelligent fuzzy rule-based e-learning recommendation system for dynamic user interests[J]. Journal of supercomputing. 2019,75(8):5145－5160.
[20] XU C, DING A, ZHAO K. A novel POI recommendation method based on trust relationship and spatial-temporal factors[J]. Electronic commerce research & applications. 2021(48),101060:1－10.
[21] 邱均平,余厚强.跨学科发文视角下我国图书情报学跨学科研究态势分析[J].情报理论与实践, 2013, 36(5): 5－10.
[22] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled-LDA: a supervised topic model for credit attribution in multi-labeled corpora [C]//Proceedings of the 2009 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2009: 248－256.
[23] 杨春艳,潘有能,赵莉.基于语义和引用加权的文献主题提取研究[J].图书情报工作, 2016,60(9):131－138.
[24] 吴小兰,章成志. 基于社交媒体的高影响力跨学科用户发现研究[J].情报学报, 2017, 36(6): 618－627.
[25] 王占,林岩.基于信任与用户兴趣变化的协同过滤方法研究[J].情报学报,2017,36(2):197－205.
[26] 刘凯玉.基于合作与引文网络的学者学术影响力评价研究[D].济南:山东师范大学,2019.
[27] 王玉斌.基于信息内容时效性改进推荐算法的策略研究与实现[D].北京:北京邮电大学,2013.
[28] 高慧颖,魏甜,刘嘉唯.基于用户聚类与动态交互信任关系的好友推荐方法研究[J]. 数据分析与知识发现, 2019, 3(10): 66－77.
[29] 刘凯玉.基于合作与引文网络的学者学术影响力评价研究[D].济南:山东师范大学,2019.
[30] 董伟,陶金虎.融合 PageRank 与评论情感倾向的在线健康社区用户影响力研究[J].图书情报工作,2021,65(11):1－10.
[31] 曾文,桂婕,徐红姣,等.基于领域的科技文献重要度评估方法研究[J].情报理论与实践,2015,38(12):73－76.
[32] 赖院根,王星.面向检索排序的论文重要度测算[J].情报理论与实践,2009,32(10):78－81.
[33] 熊回香,孟璇,叶佳鑫.基于关键词语义类型和文献老化的学术论文推荐[J].现代情报,2021,41(1):13－23.
[34] 高慧颖,魏甜,刘嘉唯.基于用户聚类与动态交互信任关系的好友推荐方法研究[J].数据分析与知识发现,2019,3(10): 66－77.
[35] 张颖怡,章成志,陈果.学术博客用户的博文分类行为研究——以科学网博客为例[J].情报学报,2016,35(11):1223－1232.
[36] 吴昊,刘东苏.社交网络中的好友推荐方法研究[J].现代图书情报技术,2015,254(1):59－65.
[37] 胡文江,胡大伟,高永兵,等.基于关联规则与标签的好友推荐算法[J].计算机工程与科学, 2013, 35(2): 109－113.

作者贡献说明:

顾佳云:提出研究方案、数据采集与处理、论文撰写与修订;

熊回香:提出研究方向和研究思路;

肖兵:论文修改。

Research on Scholar Recommendation Integrating Users' Dynamic Interests and Social Relationships in Virtual Academic Communities

Gu Jiayun Xiong Huixiang Xiao Bing
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] Considering the dynamic changes of users' interests and social relationships, this paper proposes a scholar recommendation model integrating users' dynamic interests and social relationships. [Method/Process] Firstly, using the periodical literature of different disciplines as the classified corpus, the discipline domain of scholars' blog posts was distinguished based on the labeled LDA model. Then KNN algorithm was used to classify blogs by discipline. At the same time, the change rate of subject interests was used to improve the time factor, and the dynamic interest similarity of scholars was calculated. The PageRank of scholars was calculated by using the quantitative relationship of links between scholars, and the global trust level was calculated by combining the PageRank and time value of blogs sent by scholars. Time weight was introduced into scholars' comments and recommendation interaction behaviors to calculate scholars' interactive trust level. The dynamic social trust level of scholars was obtained by integrating the global trust level and interactive trust level. Finally, the similarity of interest and trust were combined to recommend scholars. [Result/Conclusion] The scholar recommendation model integrating users' dynamic interests and social relationships in the virtual academic community can effectively improve the quality of scholar recommendation from the perspectives of dynamic interests and dynamic social relationships.

Keywords: virtual academic community dynamic interests social connections scholar recommendation Labeled-LDA topic model